Editor: Khaled El Emam, kelemam@cheo.on.ca

Seven Ways to Evaluate the Utility of Synthetic Data

Khaled El Emam | Children's Hospital of Eastern Ontario Research Institute

Access to individual-level health data is going to be critical for managing the COVID-19 pandemic and enabling society to return to some form of (new) normal functioning. Broader data access is already starting to happen. At the same time, there has been growing alarm by the privacy community about the extent and manner of the level of data sharing that is going on with such sensitive information. In South Korea, broad data sharing has already resulted in some patients being reidentified and experiencing judgment and ridicule,[1,2] and some governments have begun to reduce the amount of information being shared about COVID-19 cases.[3–8] Data synthesis can provide a solution by enabling access to useful information while ensuring reasonable privacy protections.

There are already large-scale data-sharing efforts using synthetic data. For example, tabulations from the 2020 United States Census will be based on synthetic data. Public Health England has made a large cancer registry publicly available for analysts (the Simulacrum). Additional synthesis efforts are in the works by the National Institutes of Health (NIH) and NIH-funded projects.

Synthetic health data are generated from a model that is fit to a real data set as illustrated in Figure 1. Statistical machine learning and deep learning methods are typically used to fit this model. No specific advance knowledge of how the data will be used or analyzed is required to generate useful synthetic data. Once the model is fit, it is used to generate new data from that model. The generation is stochastic; therefore, a different data set is generated from the model each time.

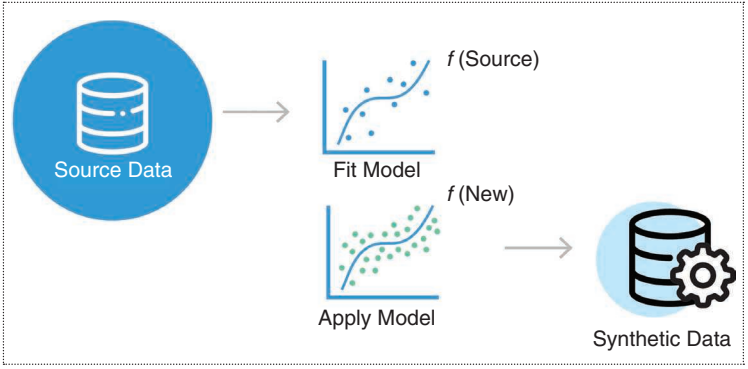

Figure 1. The basic workflow for data synthesis.

For data scientists to be comfortable using synthetic data, especially to build models that would influence public health and clinical decisions, there needs to be evidence demonstrating the utility of that data. In this article, we summarize the seven ways that the utility of synthetic data has been assessed thus far, and we close with some recommendations on their application.

Utility Assessment Methods

The following are seven methods for assessing the utility of synthetic data. In these descriptions, we will refer to the real data as the source and the synthetic data as the generated data set. The assumption is made that the objective is to make individual-level patient data broadly available, as opposed to, for example, releasing aggregate statistics or summary tables.

Utility assessment is performed by the entity performing the data synthesis before making the data available more broadly. Typically, the results of the utility assessments are documented and shared with the data users.

Replication of Studies

The default approach to assess utility is to perform an analysis on the real data and then replicate that on the synthetic data. If the same conclusions are drawn from the two different analyses, then the synthetic data are deemed to have high utility. The analysis that is chosen must be

Digital Object Identifier 10.1109/MSEC.2020.2992821
Date of current version: 9 July 2020

 Copublished by the IEEE Computer and Reliability Societies

meaningful for the applications that are expected with the synthetic data. For example, if a data set is going to be used to predict survival time, then survival models from real and synthetic data would be compared.

Another way this approach has been operationalized is to find an already-published study using the same data set and then replicate the results of that published study using the synthetic data. If the replication is successful, then that demonstrates the utility of the synthetic data set.

In practice, this is quite a convincing way to demonstrate utility. The main drawback of this approach is the need to get access to expertise in the domain in order to perform a meaningful replication analysis and interpret the results. For example, a biostatistician would typically perform these replications of health studies. Furthermore, replicating a single analysis, however meaningful, will only be partially informative about other possible analyses from the data.

Subjective Assessment by Domain Experts

A good test of data utility is whether domain experts can tell the difference between real and synthetic records. This can be evaluated, for example, by having clinicians examine a subset of records and then classify them as real or synthetic based on how realistic and plausible the data look. Standard classification accuracy metrics can be used to evaluate their performance (such as the F-score or the area under the receiver operating characteristic curve). Poor classification accuracy means that they cannot tell the difference between real and synthetic data.

Similar to replication, subjective assessments require access to those knowledgeable in the domain, such as clinicians, and they need to classify a nontrivial number of records to have stable results (for example, 100 records). This classification can be more difficult to do if the records pertain to a complex clinical process.

General Utility Metrics

The most commonly used approach to the evaluation of synthetic data is to use generic metrics. For example, one can compute the distance between the real variable distributions and the synthetic distributions, or one can compare the correlations among the variables in the synthetic data and the real data. These types of metrics do not consider the specific analyses that would eventually be performed with the data. Rather, they assess general statistical parameters and model evaluation results for plausible classes of analyses that would be performed on the data.

For data scientists to be comfortable using synthetic data, especially to build models that would influence public health and clinical decisions, there needs to be evidence demonstrating the utility of that data.

It is also possible to assess the distinguishability of the synthetic data. This involves building a classification model that can distinguish between real and synthetic data. If the model is not able to distinguish between them, then the utility is high. It is an automated version of the subjective assessment approach mentioned in the section "Subjective Assessment by Domain Experts."

General metrics have the advantage of being largely automated and can provide a good perspective on the utility of the data. If the data set is not deemed adequate on the general metrics, then it will likely not perform well on any of the other tests.

Bias and Stability Assessment

Because data synthesis is stochastic, a different set of values is produced each time a synthetic data set is generated from the fitted model. One approach that has been used to determine synthetic data bias is to generate a large number of synthetic data sets and then compute the general utility metrics evaluation on the average. The variation in these parameters has also been evaluated to determine the stability of the synthetic data.

When replicating studies, bias and stability can also be evaluated on the replications. This is another way to determine the reliability of the replication results. This is an interesting utility assessment to perform from an overall statistical perspective. If the metrics computed from synthetic data are biased in a systematic way or have nontrivial variability, then the fitted generative model's behavior cannot be relied on.

Structural Similarity

Although this would seem like a minor detail, in practice, it is a very important way to evaluate synthetic data utility given the common ways that synthetic data are used. Structural similarity means that the synthetic data should pass edit checks and have the same variable types and formats, variable names, metadata, and file formats, as well table names and structure. This allows analysts to use the same analysis code to analyze the synthetic data as the real data.

A common use case for synthetic data is to test statistical programs; being able to run the same code makes the synthetic data very useful from a practical standpoint. Another use case is to validate the results from the synthetic data on the real data using a validation server, as illustrated in Figure 2.

This allows analysts to run their code on real data without accessing it. The returned results are manually reviewed for disclosure risks, and therefore validation on real data is performed once at the end of the analysis. For this to work, the code must run without modification on both types of data.

Comparison With Public Aggregate Data

There is an increasing amount of aggregate data shared publicly on COVID-19 cases, mortality, comorbidities, and concomitant medications. An assessment of synthetic data can compare the computation of these statistics from the synthetic data with the public results to see if they are congruent. The utility of the synthetic data set would be deemed higher if it is telling the same story as the public data.

Comparison With Other Privacy-Enhancing Technologies

To enable broad analytic access to COVID-19 data, a number of different approaches can be used, each with its own strengths and weaknesses, such as pseudonymization, deidentification, federated analysis, and protocols based on homomorphic encryption. All of the assessments can be performed on these other methods as well and then compared to synthetic data.

This type of assessment can only inform us of the relative utility of data synthesis to other techniques that can be applied. Arguably, this could be one factor to consider when choosing an approach for providing data access.

The seven approaches described in this article have been used in practice and in the literature to evaluate the utility of synthetic data. To enable broad access to individual-level health data, it is not only important to protect patient privacy, but also to ensure that the utility of that data, after any transformations, is still sufficiently high for meaningful analysis.

In Table 1 we prioritize the approaches in terms of how useful they are in ensuring that the

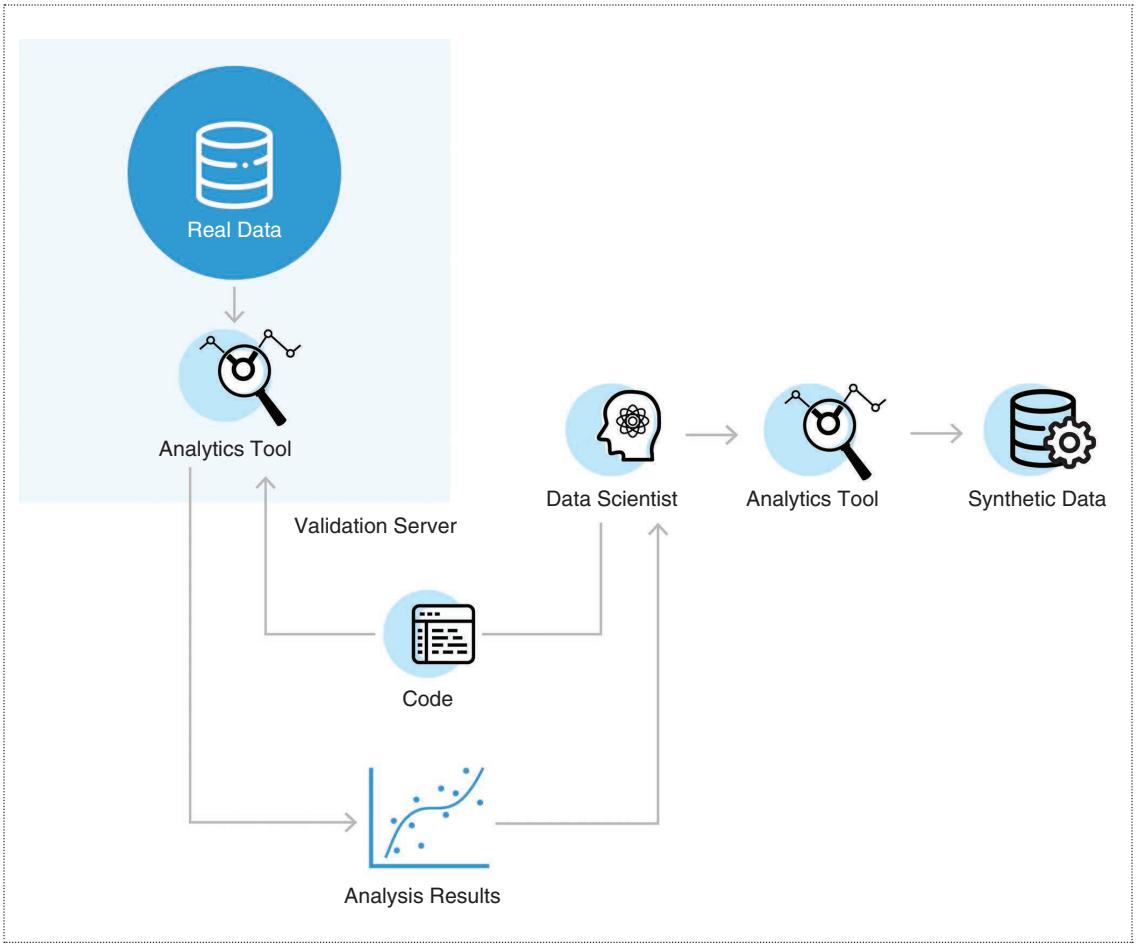

Figure 2. The setup for a validation server.

Table 1. Applicability of different utility assessment methods for synthetic data.

Utility assessment approach	Explanatory comments	Applicability
Structural similarity	This is critical. If the data is not structurally similar, then that just makes it harder for analysts to use it.	Perform for every data set
General utility metrics	This is critical. Every data set needs to pass a minimal set of utility metrics. This is relatively easy to do because it can be largely automated.	Perform for every data set
Replication of studies	Replication is a convincing way to demonstrate that a synthetic data method can be relied upon. It is a time-consuming process that requires domain expertise.	Evaluate methodology
Subjective assessment by domain experts	This type of assessment of the synthesis methodology can also be quite convincing. It is a more challenging assessment to perform.	Evaluate methodology
Bias and stability assessment	This is a generally useful type of assessment to perform for every synthetic data release. However, the weight of evidence it adds to the utility of a synthetic data set is smaller than the other approaches.	Every data set
Comparison with public aggregate data	When available, comparisons to public data will enhance confidence in a synthesis methodology.	Methodology evaluation
Comparison with other PETs	This type of assessment is useful to perform at some point to help decision makers decide the relative strengths and weaknesses of particular PETs for providing data access.	Methodology evaluation

PETs: privacy-enhancing technologies.

synthetic data will be useful for analysis purposes. We make a distinction between a utility analysis that must be performed every time a synthetic data set is generated versus an analysis that should be performed on the synthesis methodology. For the latter, a utility assessment demonstrates that the synthesis approach works well in practice but would be too difficult to perform for every single data set.

Note that we did not discuss the privacy of data synthesis. Our assumption is that the synthesis models were not overfit and that the identity disclosure risks from the synthetic data are very small. We limited our narrative to utility. Of course, these utility assessment approaches are applicable to other methods that can be used to protect the privacy of individuals by enabling access to individual-level nonpersonal data. ■

References

1. "Coronavirus privacy: Are South Korea's alerts too revealing?" *BBC*, London, Mar. 5, 2020. [Online]. Available: https://bbc.in/2Z7IPCL
2. N. Kim, "'More scary than coronavirus': South Korea's health alerts expose private lives," *The Guardian*, Mar. 6, 2020. [Online]. Available: https://bit.ly/3fWFM6D
3. R. Rocha, "The data-driven pandemic: Information sharing with COVID-19 is 'unprecedented,'" *CBC News*, Canada, Mar. 17, 2020. [Online]. Available: https://bit.ly/3bDxhtu
4. K. Rackley, "DHEC, state authorities address privacy issues, information about coronavirus case specifics," *Aiken Standard*, Aiken, SC, Apr. 4, 2020. [Online]. Available: https://bit.ly/2Tbm8Kp
5. J. Hinkle, "Framingham one of several cities and towns told by DPH to limit information about residents who test positive for coronavirus," *Wicked Local*, Mar. 28, 2020. [Online]. Available: https://bit.ly/2WZtXny
6. A. McCallum, "Janesville and Rock County officials clash over sharing of COVID-19 information," *GazetteXtra*, Janesville, WI, Apr. 5, 2020. [Online]. Available: https://bit.ly/3fXmjCw
7. L. Hancock, "Ohio health director cites privacy concerns as local health departments withhold coronavirus details," Cleveland, OH, Apr. 3, 2020. [Online]. Available: https://bit.ly/2z5RkDV
8. K. Hill, "Spokane health officials providing more information about COVID-19 patients, but it remains unclear where they're being treated," *The Spokesman-Review*, Spokane, WA. Apr. 6, 2020. [Online]. Available: https://bit.ly/2WCuFrV

Khaled El Emam is a senior scientist at the Children's Hospital of Eastern Ontario Research Institute, a professor in the Faculty of Medicine at the University of Ottawa, and cofounder and director at Replica Analytics Ltd. Contact him at kelemam@uottawa.ca.